Progress with MPI work

Leiqing Cai

10/25/2013

Overall goal/Process

- To contribute to the intra-datacenter networking research effort
- Improve latency of MPI communications in CESM applications

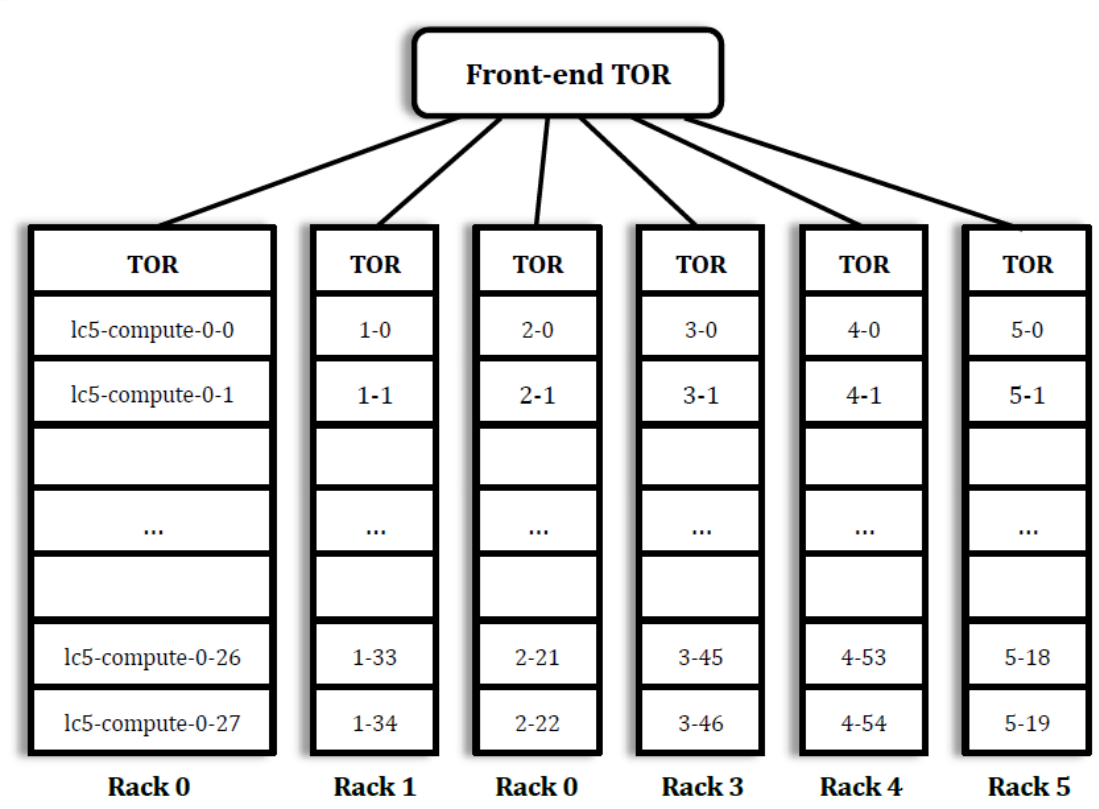
UNIVERSITY of VIRGINIA

Starting Steps

- Started with the given mpi_pingpong.c file, a benchmark program for MPI communication
- Wrote a ping script to figure out the network topology of Fir (the Linux cluster at UVA)
- Wrote PBS scripts to run mpi_pingpong on the cluster
- Collected measurements
- Wrote a data processing program (in C++) to analyze the results of the experiments

Fir

- A Linux-based cluster managed by UVACSE (UVa Alliance for Computational Science & Engineering) and operated by ITS (Information Technology Services) at University of Virginia
- Hosts are interconnected by Ethernet switches



- 6 Racks
- 208 Nodes
- TOR = Top of Rack

Results

- 2 Rounds of experiments
  - First round: 40 jobs. The two specified hosts are on the same rack.
  - Second round: 40 jobs, The two specified hosts are on different racks.
- Parameters for each job
  - 5000 iterations for each communicating pair to reduce measurement error
  - Message size: 1024 Bytes (one packet)
  - Non-persistent MPI communication: MPI_Isend() and MPI_Irecv()

Table 1 mpi_pingpong experiment results (Unit of time in μs)

	2-hop pairs	**4-hop pairs**
Average Delay	62.608	71.595
Standard Deviation	16.641	19.126
Min	48.094	48.516
Max	95.714	116.49

Demo

Lessons learned

- More control over PBS jobs
  - Request for particular hosts in a cluster to run PBS jobs
  - Request for multiple cores on a host to run PBS jobs
- Non-blocking non-persistent communications and persistent communications with MPI
- Unix shell scripts with complex syntax
- Network topology on Fir

UNIVERSITY of VIRGINIA

Next Steps

- Run mpi_pingpong on JellyStone
  - Study the network topology on JellyStone
  - Learn InfiniBand architecture and protocols
  - Collect and analyze larger data sets
  - Compare results between persistent and non-persistent MPI communications
- Obtain measurements by using ibdump
- Run MPI experiments on Denali, a cluster at University of New Mexico (PRObE): InfiniBand

UNIVERSITY of VIRGINIA